BOOKMATES:
un recommendation system per compagni di lettura

INDICE

1. ABSTRACT
2. STATO DELL'ARTE
3. INTRODUZIONE
4. DATI E STRUMENTI UTILIZZATI
5. LAVORO SVOLTO
    - 5.1 SISTEMA DI RACCOMANDAZIONE
    - 5.2 SENTIMENT ANALYSIS E XAI
    - 5.3 INTERFACCIA GRADIO
6. LIMITI E PROSPETTIVE FUTURE

1. ABSTRACT

In questo progetto abbiamo utilizzato due dataset: il primo dataset contiene informazioni dettagliate sui libri, tra cui titolo, autore, prezzo, data di uscita e descrizione. Il secondo dataset, invece, contiene tutte le recensioni degli utenti relative ai libri presenti nel primo dataset. L'obiettivo principale del progetto è stato la creazione di un sistema di raccomandazione di libri.

Per sviluppare questo sistema, abbiamo impiegato alcune tecniche avanzate di elaborazione del linguaggio naturale (NLP) come i word embedding (ottenuti utilizzando SBert), la cosine similarity e il TF-IDF (Term Frequency-Inverse Document Frequency). Il sistema (che sfrutta un'interfaccia grafica sviluppata con Gradio), una volta ricevuto in input un libro da parte dell'utente (e trovato con la funzione Rapidfuzz il libro più simile, in modo da evitare eventuali errori di battitura dell'utente in fase di inserimento del libro), è in grado di identificare e stampare i 10 libri più simili basandosi su una combinazione delle due tecniche sopra descritte (word embedding e TF-IDF) e sul punteggio medio assegnato dagli utenti nelle recensioni. Inoltre, abbiamo effettuato una sentiment analysis utilizzando BERT per ottenere i token delle parole e una CNN (Convolutional Neural Network) per classificare le recensioni. Al fine di comprendere i motivi per cui un libro è piaciuto o meno agli utenti, è stato utilizzato LIME (Local Interpretable Model-agnostic Explanations).

Infine, i risultati di LIME sono stati inseriti nell'output del sistema di raccomandazione per mostrare agli utenti il motivo delle recensioni positive dei libri che gli vengono consigliati. Nell'output di Gradio abbiamo mostrato anche un breve riassunto del libro, utile agli utenti per decidere quale libro scegliere, ottenuto con il modello BART.

2. STATO DELL'ARTE e LETTERATURA

Al giorno d'oggi, i sistemi di raccomandazione per i libri si trovano in una fase avanzata di sviluppo e implementazione; grazie agli enormi progressi nell'analisi dei dati e nell'intelligenza artificiale, questi sistemi sono in grado di offrire raccomandazioni personalizzate e precise agli utenti.

Utilizzando algoritmi sofisticati come collaborative filtering, content-based filtering, e hybrid approaches, i sistemi di raccomandazione possono considerare non solo le preferenze passate degli utenti ma anche caratteristiche dettagliate dei libri stessi, come il genere, l'autore, il contenuto e le recensioni degli utenti.
Inoltre, l'integrazione di tecniche di machine learning e deep learning ha permesso una maggiore capacità di predizione e adattamento alle preferenze individuali degli utenti, rendendo i sistemi di raccomandazione per i libri uno strumento essenziale per migliorare l'esperienza di lettura e promuovere la scoperta di nuovi libri.

Alcuni esempi, già presenti sul mercato, di sistemi di raccomandazione per libri sono: Goodreads (che utilizza un approccio basato sul collaborative filtering) e Amazon (che utilizza un sistema ibrido, basato su collaborative filtering, content-based filtering e knowledge-based).

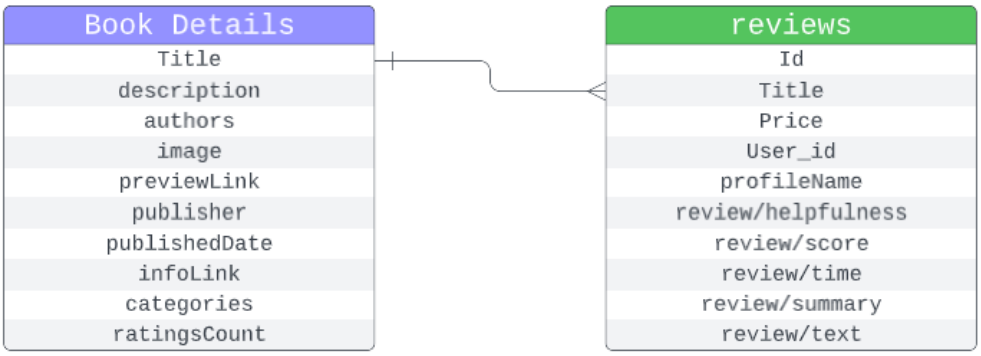
3. INTRODUZIONE

Il crescente utilizzo di sistemi di raccomandazione ha rivoluzionato il modo in cui consumiamo contenuti digitali, con applicazioni ben consolidate per canzoni, film e serie TV.
In un'epoca in cui la tecnologia pervade ogni aspetto della nostra vita, nel nostro progetto abbiamo voluto unire un oggetto fisico e presente da millenni, come il libro, con le tecnologie più avanzate nel campo dell'analisi dei dati e dell'intelligenza artificiale.

Questo progetto mira a creare un sistema di raccomandazione per libri, al fine di migliorare l'esperienza di scoperta e fruizione dei libri, offrendo suggerimenti più mirati e accurati, basandosi sia sulla somiglianza della sinossi dei libri già letti sia sulle recensioni dei lettori.

4. DATI E STRUMENTI UTILIZZATI

Per questo progetto abbiamo utilizzato due dataset provenienti da Kaggle, ottenuti mediante scraping da Amazon. Il primo dataset contiene informazioni dettagliate sui libri, tra cui titolo, descrizione, autore, editore e altre caratteristiche rilevanti. Il secondo dataset include tutte le recensioni degli utenti relative ai libri presenti nel primo dataset, le variabili più importanti sono: l'id dell'utente, il titolo, il prezzo, il punteggio assegnato su scala discreta (da 1 a 5) dall'utente e il testo della recensione scritta dall'utente.
Il primo dataset conteneva circa 212 mila libri univoci, poi ridotti a 32 mila dopo l'eliminazione dei record con valori mancanti e dei libri con meno di 10 recensioni per ciascuno; mentre il secondo dataset conteneva circa 3 milioni di recensioni degli utenti, poi ridotte a 500 mila sia per l'eliminazione dei record con valori mancanti nelle variabili di interesse sia per motivi pratici e computazionali.
Per effettuare la join tra i due dataset è stata utilizzata come chiave univoca la variabile *Title*.



I dati sopra descritti sono stati elaborati utilizzando diverse tecniche avanzate di elaborazione del linguaggio naturale (NLP).
In particolare, sono stati utilizzati i word embedding (rappresentazioni vettoriali) delle descrizioni dei libri, tramite il modello SBERT *paraphrase-MiniLM-L6-v2* (abbiamo utilizzato la versione presente sul sito HuggingFace). In questo modo abbiamo ottenuto la rappresentazione delle parole (presenti nelle descrizioni dei libri) in uno spazio vettoriale (matrice degli embeddings), catturando le relazioni semantiche tra i termini; infine, si è calcolata la cosine similarity tra i singoli embeddings.
Inoltre, abbiamo utilizzato anche l'algoritmo TF-IDF (Term Frequency-Inverse Document Frequency), esso calcola l'importanza di ogni termine in ciascuna descrizione di ogni libro, tenendo conto della frequenza dei termini nei documenti.
Per entrambi gli algoritmi è stata utilizzata la cosine similarity al fine di misurare la somiglianza tra le rappresentazioni vettoriali dei libri.

Successivaente, abbiamo svolto una sentiment analysis delle recensioni utilizzando BERT (Bidirectional Encoder Representations from Transformers) per ottenere i token delle parole e una CNN (Convolutional Neural Network) per classificare il sentiment delle recensioni.
Al fine di rendere il modello più interpretabile e comprendere le ragioni dietro le preferenze

degli utenti, è stato utilizzato l'algoritmo di explainable-AI LIME (Local Interpretable Model-agnostic Explanations), il quale ci ha mostrato quali sono state le parole più influenti che hanno portato gli utenti a recensire positivamente o negativamente i libri.

Infine, abbiamo utilizzato un modello pre-addestrato BART (Bidirectional and Auto-Regressive Transformers) sviluppato da Facebook AI, per ottenere un riassunto della descrizione del libro da mostrare all'utente nella finestra di Gradio, per aiutarlo nella scelta del prossimo libro.

5.1 SISTEMA DI RACCOMANDAZIONE

Di seguito una rapida spiegazione del funzionamento del sistema.

1. Input dell'utente: l'utente fornisce il titolo di un libro; per ovviare ad eventuali errori di battitura dell'utente abbiamo implementato due strategie:
    - Convertire in minuscolo sia l'input dell'utente sia tutti i libri presenti nel nostro dataset;
    - Utilizzare la funzione Rapidfuzz, già presente in Python, che restituisce la corrispondenza più simile alla stringa di input dell'utente all'interno di un elenco di stringhe, utilizzando un algoritmo di confronto delle stringhe basato sulla distanza di Levenshtein.

2. Matching del titolo: utilizzando i due strumenti descritti sopra, il sistema trova il titolo più simile nel dataset.

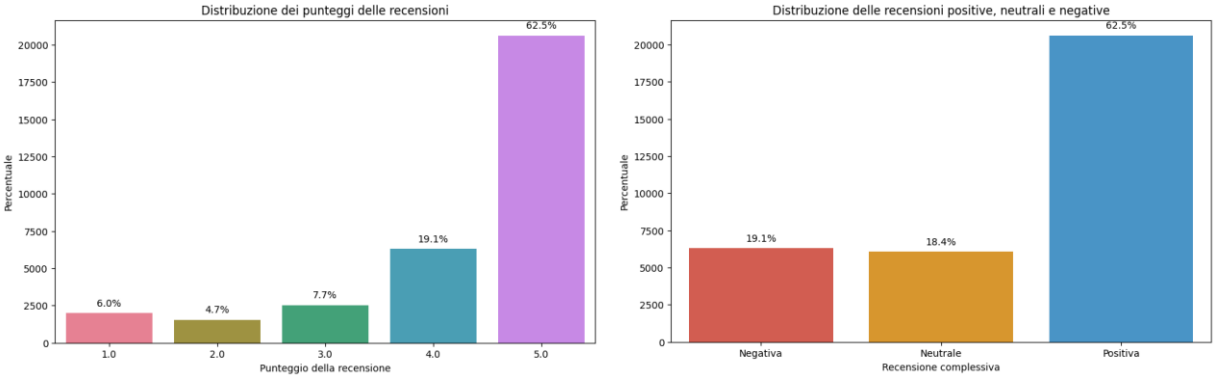
3. Calcolo delle similarità:
    - Per gli embeddings, vengono cercati (e ordinati in ordine decrescente) i valori di cosine similarity (già calcolata in precedenza, in modo da rendere più rapido il funzionamento del sistema di raccomandazione) tra il libro fornito dall'utente e tutti gli altri libri.
    - Per TF-IDF, viene utilizzata la matrice TF-IDF ridotta per calcolare la cosine similarity e si procede come per gli embeddings.

4. Output delle raccomandazioni (tramite finestra grafica di Gradio): il sistema restituisce due liste di libri consigliati, una lista basata sulla similarità degli embeddings e l'altra basata sulla similarità ottenuta dal TF-IDF.
   Inoltre, nella finestra grafica di Gradio, per fornire all'utente una rapida idea del contenuto del libro, abbiamo aggiunto una colonna che mostra il riassunto della descrizione del libro, utilizzando BART (Bidirectional and Auto-Regressive Transformers), un modello pre-addestrato sviluppato da Facebook AI.

5. Feedback: ogni utente può selezionare quale delle due liste di libri consigliati preferisce, il sistema si salva questa preferenza al fine di avere (a tendere) un dataset che contiene tutti i feedback degli utenti, da utilizzare come reinforcement learning per il sistema di raccomandazione.

5.2 SENTIMENT ANALYSIS E XAI

Si è voluto analizzare anche le recensioni dei libri, eseguendo una sentiment analysis utilizzando un modello di deep learning basato sulle reti neurali convoluzionali. Nel dataset originale il punteggio delle recensioni variava da 1 a 5 (con un forte sbilanciamento per la classe “5”, lo score più alto); si è scelto di dividere le recensioni in 3 classi (negativa: livello 1, 2 e 3, neutrale: livello 4 e positiva: livello 5), si è optato per questa categorizzazione per non creare classi ancora più sbilanciate. Infine, la classe “neutrale” è stata eliminata e la sentiment analysis è stata svolta su tutte le recensioni “negative” e “positive”.



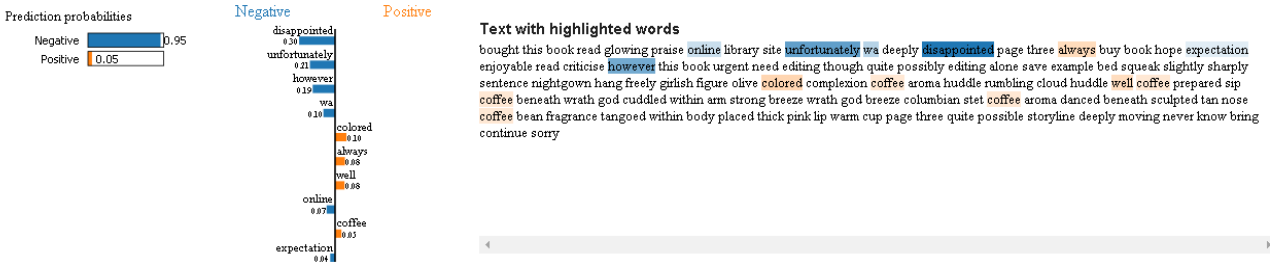
Prima di tutto, si è svolto il pre-processing su tutti i testi: si sono eliminate le stopwords, abbiamo fatto la lemmatization, rimosso caratteri ininfluenti (come caratteri speciali e la punteggiatura) e abbiamo creato i bigrams e gli ngrams.
I testi puliti sono stati poi tokenizzate con BERT, suddivisi in dataset di test e di train ed infine si è addestrato un modello di reti neurali convoluzionali (CNN).
Dopo due epoche, il modello ha ottenuto un accuracy pari all’87% nella categorizzazione delle recensioni in positive e negative.

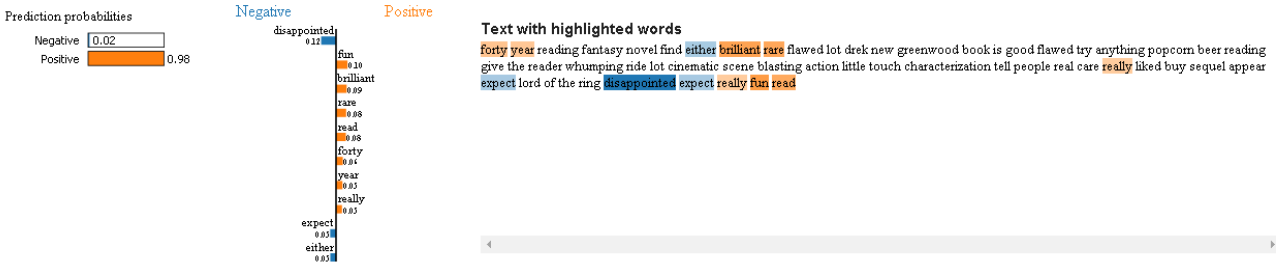
```
Epoch 1/2
1335/1335 [==============================] - 342s 255ms/step - loss: 0.3752 - accuracy: 0.8367
Epoch 2/2
1335/1335 [==============================] - 343s 257ms/step - loss: 0.1780 - accuracy: 0.9348
334/334 [==============================] - 18s 52ms/step - loss: 0.3125 - accuracy: 0.8689
334/334 [==============================] - 18s 53ms/step
              precision    recall  f1-score   support

           0       0.71      0.72      0.71      1216
           1       0.92      0.91      0.91      4123

    accuracy                           0.87      5339
   macro avg       0.81      0.82      0.81      5339
weighted avg       0.87      0.87      0.87      5339
```

I risultati della classificazione, ottenuta con il modello di reti neurali convoluzionali, sono stati “inseriti” nell’algoritmo di explainable-AI Lime, il quale ha evidenziato le parole principali che hanno portato a categorizzare una recensione di un libro come positiva o negativa.
Le parole principali evidenziate da Lime, infine, sono state inserite all’interno delle informazioni mostrate all’utente, con lo scopo di spiegargli il motivo del punteggio medio delle recensioni.



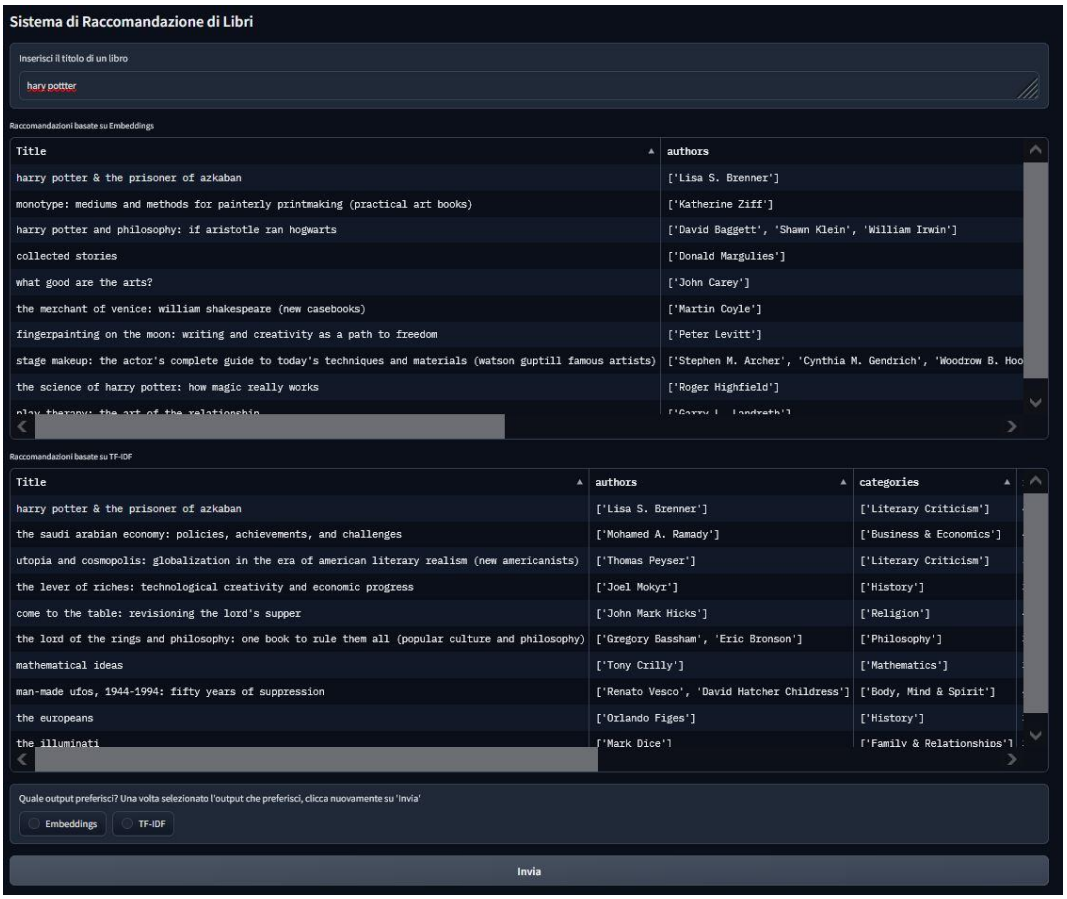


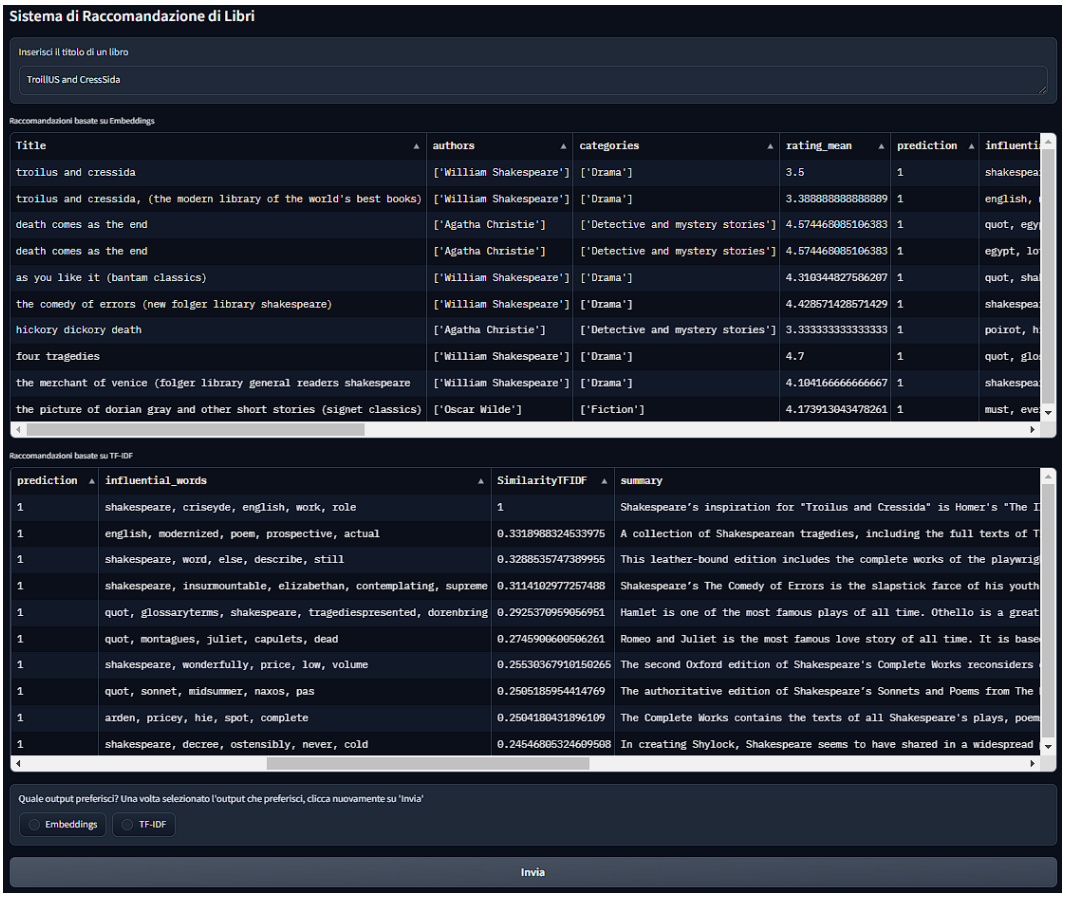
5.3 GRADIO

L'interfaccia utente è costruita utilizzando il framework Gradio e contiene i seguenti elementi:

- Un campo di testo dove l'utente inserisce il titolo dell'ultimo libro che ha letto e che ha apprezzato;
- Due tabelle che mostrano le due liste dei libri consigliati (con le informazioni principali sul libro, le parole più rilevanti nelle recensioni selezionate dall'algoritmo di xai Lime, il riassunto della descrizione del libro e il punteggio di similarità) dal sistema di raccomandazione basato sulle due tecniche (embeddings e TF-IDF);
- Un pulsante di feedback per selezionare il metodo di raccomandazione che ha fornito i risultati che l'utente ha gradito maggiormente;
- Un pulsante di invio per generare le raccomandazioni (dopo aver inserito il testo) e per inviare i feedback (dopo aver scelto quale delle due liste preferisce).

Le funzioni (visibili nella seconda schermata di seguito) che mostrano la variabile "prediction" sulla positività o negatività della recensione, le parole più influenti (evidenziate da Lime) e il riassunto della descrizione del libro, sono state implementate per un piccolo set di righe per problemi computazionali.


Sistema di Raccomandazione di Libri
Inserisci il titolo di un libro
hary pottter
Raccomandazioni basate su Embeddings
Title
authors
harry potter & the prisoner of azkaban
['Lisa S. Brenner']
monotype: mediums and methods for painterly printmaking (practical art books)
['Katherine Ziff']
harry potter and philosophy: if aristotle ran hogwarts
['David Baggett', 'Shawn Klein', 'William Irwin']
collected stories
['Donald Margulies']
what good are the arts?
['John Carey']
the merchant of venice: william shakespeare (new casebooks)
['Martin Coyle']
fingerpainting on the moon: writing and creativity as a path to freedom
['Peter Levitt']
stage makeup: the actor's complete guide to today's techniques and materials (watson guptill famous artists)
['Stephen M. Archer', 'Cynthia M. Gendrich', 'Woodrow B. Hoo
the science of harry potter: how magic really works
['Roger Highfield']
Raccomandazioni basate su TF-IDF
Title
authors
categories
harry potter & the prisoner of azkaban
['Lisa S. Brenner']
['Literary Criticism']
the saudi arabian economy: policies, achievements, and challenges
['Mohamed A. Ramady']
['Business & Economics']
utopia and cosmopolis: globalization in the era of american literary realism (new americanists)
['Thomas Peyser']
['Literary Criticism']
the lever of riches: technological creativity and economic progress
['Joel Mokyr']
['History']
come to the table: revisioning the lord's supper
['John Mark Hicks']
['Religion']
the lord of the rings and philosophy: one book to rule them all (popular culture and philosophy)
['Gregory Bassham', 'Eric Bronson']
['Philosophy']
mathematical ideas
['Tony Crilly']
['Mathematics']
man-made ufos, 1944-1994: fifty years of suppression
['Renato Vesco', 'David Hatcher Childress']
['Body, Mind & Spirit']
the europeans
['Orlando Figes']
['History']
Quale output preferisci? Una volta selezionato l'output che preferisci, clicca nuovamente su 'Invia'
Embeddings
TF-IDF
Invia


Sistema di Raccomandazione di Libri
Inserisci il titolo di un libro
TroilUS and CressSida
Raccomandazioni basate su Embeddings
Title
authors
categories
rating_mean
prediction
influenti
troilus and cressida
['William Shakespeare']
['Drama']
3.5
1
shakespea
troilus and cressida, (the modern library of the world's best books)
['William Shakespeare']
['Drama']
3.388888888888889
1
english,
death comes as the end
['Agatha Christie']
['Detective and mystery stories']
4.574468085106383
1
quot, egy
death comes as the end
['Agatha Christie']
['Detective and mystery stories']
4.574468085106383
1
egypt, lo
as you like it (bantam classics)
['William Shakespeare']
['Drama']
4.310344827586207
1
quot, sha
the comedy of errors (new folger library shakespeare)
['William Shakespeare']
['Drama']
4.428571428571429
1
shakespea
hickory dickory death
['Agatha Christie']
['Detective and mystery stories']
3.333333333333333
1
poirot, h
four tragedies
['William Shakespeare']
['Drama']
4.7
1
quot, glo
the merchant of venice (folger library general readers shakespeare
['William Shakespeare']
['Drama']
4.104166666666667
1
shakespea
the picture of dorian gray and other short stories (signet classics)
['Oscar Wilde']
['Fiction']
4.173913043478261
1
must, eve
Raccomandazioni basate su TF-IDF
prediction
influential_words
SimilarityTFIDF
summary
1
shakespeare, criseyde, english, work, role
1
Shakespeare's inspiration for "Troilus and Cressida" is Homer's "The I
1
english, modernized, poem, prospective, actual
0.3318988324533975
A collection of Shakespearean tragedies, including the full texts of T
1
shakespeare, word, else, describe, still
0.3288535747389955
This leather-bound edition includes the complete works of the playwrig
1
shakespeare, insurmountable, elizabethan, contemplating, supreme
0.3114102977257488
Shakespeare's The Comedy of Errors is the slapstick farce of his youth
1
quot, glossaryterms, shakespeare, tragediespresented, dorenbring
0.2925370959056951
Hamlet is one of the most famous plays of all time. Othello is a great
1
quot, montagues, juliet, capulets, dead
0.2745900600506261
Romeo and Juliet is the most famous love story of all time. It is base
1
shakespeare, wonderfully, price, low, volume
0.2553036791015026 5
The second Oxford edition of Shakespeare's Complete Works reconsiders
1
quot, sonnet, midsummer, naxos, pas
0.2505185954414769
The authoritative edition of Shakespeare's Sonnets and Poems from The
1
arden, pricey, hie, spot, complete
0.2504180431896109
The Complete Works contains the texts of all Shakespeare's plays, poem
1
shakespeare, decree, ostensibly, never, cold
0.24546805324609508
In creating Shylock, Shakespeare seems to have shared in a widespread
Quale output preferisci? Una volta selezionato l'output che preferisci, clicca nuovamente su 'Invia'
Embeddings
TF-IDF
Invia

5. LIMITI e PROSPETTIVE FUTURE

Nonostante il nostro sistema di raccomandazione di libri abbia mostrato risultati promettenti, ci sono alcuni limiti significativi da considerare:

1. **Riduzione della dimensionalità del dataset**: per velocizzare i tempi computazionali, abbiamo dovuto ridurre la dimensionalità del dataset. Questa riduzione ha comportato una perdita significativa di informazioni, limitando potenzialmente l'accuratezza e la varietà delle raccomandazioni generate dal sistema. Un dataset completo e dettagliato avrebbe potuto offrire raccomandazioni più precise e personalizzate.
2. **Utilizzo di modelli leggeri**: a causa delle limitazioni di ram di Google Colab, siamo stati obbligati a utilizzare modelli leggeri. I modelli utilizzati sono meno onerosi in termini computazionali, non possono competere in termini di complessità e capacità di apprendimento con modelli più avanzati. Questo ha limitato la performance del sistema.

Potenzialità Future

1. **Miglioramento continuo grazie ai feedback degli utenti**: il sistema migliorerà progressivamente grazie ai feedback continui forniti dagli utenti. Ogni valutazione e ogni raccomandazione accettata o rifiutata contribuirà ad affinare gli algoritmi di raccomandazione, rendendoli sempre più precisi e personalizzati.
2. **Integrazione di Neo4j per l'analisi delle reti**: con i feedback e i libri inseriti dagli utenti, si potrebbe creare una rete in Neo4j per analizzare le connessioni tra gli utenti. Questo permetterebbe di identificare quali utenti hanno gusti simili, facilitando la creazione di reti di lettori con interessi affini e migliorando ulteriormente le raccomandazioni.
3. **Espansione del dataset e utilizzo di modelli avanzati**: in futuro, con accesso a risorse computazionali più potenti, sarà possibile utilizzare dataset più completi e modelli di machine learning più avanzati. Questo potrebbe portare a un miglioramento significativo nella qualità delle raccomandazioni, permettendo al sistema di gestire una maggiore varietà di dati e di apprendere in modo più efficiente dalle interazioni degli utenti.
4. **Personalizzazione avanzata**: con l'aumento della quantità e della qualità dei dati disponibili, il sistema potrebbe implementare funzionalità di personalizzazione avanzata, come la raccomandazione di libri basata su stati d'animo, momenti della giornata o contesti specifici.